Artificial Intelligence Center

Papers matching pattern "karp"

● **Search** Any Field ▾ in Papers ▾ and display in Standard Form ▾

submit (Search Help)

P. D. **Karp**, V. K. Chaudhri, and S. M. Paley, "A Collaborative Environment for Authoring Large Knowledge Bases," *Journal of Intelligent Information Systems*, vol. 13, pp. 155--194, 1999.

Abstract: Collaborative knowledge base (KB) authoring environments are critical for the construction of high-performance KBs. Such environments must support rapid construction of KBs by a collaborative effort of teams of knowledge engineers through reuse of existing knowledge and software components. They should support the manipulation of knowledge by diverse problem-solving engines even if that knowledge is encoded in different languages and by different researchers. They should support large KBs and provide a scalable and interoperable development infrastructure. In this paper, we present an environment that satisfies many of these goals.

V. K. Chaudhri, A. Farquhar, R. Fikes, P. D. **Karp**, and J. P. Rice, "OKBC: A Programmatic Foundation for Knowledge Base Interoperability," in *Proceedings of the AAAI-98*, (Madison, WI), 1998.

Abstract: The technology for building large knowledge bases (KBs) is yet to witness a breakthrough so that a KB can be constructed by the assembly of prefabricated knowledge components. Most of the current KB development tools can only manipulate knowledge residing in the knowledge representation system (KRS) for which the tools were originally developed. Open Knowledge Base Connectivity (OKBC) is an application programming interface for accessing KRSs, and was developed to enable the construction of reusable KB tools. OKBC improves upon its predecessor, the Generic Frame Protocol (GFP), in several significant ways. In this paper, we discuss technical design issues faced in the development of OKBC, highlight how OKBC improves upon GFP, and report on practical experiences in using it.

P. **Karp**, M. Riley, S. Paley, A. Pellegrini-Toole, and M. Krummenacker, "EcoCyc: Electronic encyclopedia of E. coli genes and metabolism," *Nuc. Acids Res.*, vol. 25, no. 1, 1997.

Abstract: The Encyclopedia of E. coli Genes and Metabolism (EcoCyc) is a database that combines information about the genome and the intermediary metabolism of E. coli. It describes 2970 genes of E. coli, 547 enzymes encoded by these genes, 702 metabolic reactions that occur in E. coli, and the organization of these reactions into 107 metabolic pathways. The EcoCyc graphical user interface allows scientists to query and explore the

EcoCyc database using visualization tools such as genomic-map browsers and automatic layouts of metabolic pathways. EcoCyc spans the space from sequence to function to allow scientists to investigate an unusually broad range of questions. EcoCyc can be thought of as both an electronic review article because of its copious references to the primary literature, and as an in silicio model of E. coli metabolism that can be probed and analyzed through computational means.

V. K. Chaudhri and P. D. **Karp**. "Querying Schema Information." *Proceedings of the 4th International Workshop Knowledge Representation Meets Databases (KRDB'97)*, pp. 4--1 to 4--6, 1997.

Abstract: Schema queries can play an important role while retrieving information from multiple sources, for example, in query formulation and in query optimization. We identify four classes of schema queries that we have found useful while designing an application programming interface for frame representation systems (FRSs): taxonomic, frame structure, constraint and class comparison queries. We propose a scheme for direct support for these queries in a mediator language such as Object Query Language (OQL).

P. D. **Karp**. "A protocol for maintaining multidatabase referential integrity." in *Proceedings of the 1996 Pacific Symposium on Biocomputing*, 1996.

Abstract: The bioinformatics community is becoming increasingly reliant on the creation of links among biological databases (DBs) as a foundation for DB interoperability. For example, a link might be created from a protein in one DB (such as PIR), to a gene in another DB (such as GDB), by storing the unique identifier (id) of the gene object within an attribute of the protein object. User interfaces can then support navigation from the protein to the gene, and multiDB queries can join the protein with the gene. The unique id of the gene is serving as a foreign key. However, a variety of factors, such as changes in the underlying biology, can cause object ids to become invalid, thus producing invalid links among DBs. Invalid links are a violation of multidatabase referential integrity. We propose a network protocol whereby a database administrator can provide information about changes to the identifiers of objects in their database via Internet, to allow other databases to maintain referential integrity. We request comments from the bioinformatics community for the purpose of building a consensus on the proposed protocol.

P. **Karp**, M. Riley, S. Paley, and A. Pellegrini-Toole. "EcoCyc: Electronic encyclopedia of E. coli genes and metabolism." *Nuc. Acids Res.*, vol. 24, no. 1, pp. 32--40, 1996.

Abstract: The Encyclopedia of E. coli Genes and Metabolism (EcoCyc) is a database that combines information about the genome and the intermediary metabolism of E. coli. It describes 2034 genes of E. coli, 306 enzymes encoded by these genes, 580 metabolic reactions that occur in E. coli, and the organization of these reactions into 100 metabolic pathways. The EcoCyc graphical user interface allows scientists to query and explore the EcoCyc database using visualization tools such as genomic-map browsers and automatic layouts of metabolic pathways. EcoCyc spans the space from sequence to function to allow scientists to investigate an unusually broad range of questions. EcoCyc can be thought of as both an electronic review article because of its copious references to the primary literature, and as an in silico model of E. coli that can be probed and analyzed through computational means.

P. D. **Karp**, C. Ouzounis, and S. M. Paley. "HinCyc: A knowledge base of the complete genome and

metabolic pathways of H. influenzae," in *Proceedings of the Fourth International Conference on Intelligent Systems for Molecular Biology*, (Menlo Park, CA), AAAI Press, 1996.

Abstract: We present a methodology for predicting the metabolic pathways of an organism from its genomic sequence by reference to a knowledge base of known metabolic pathways. We applied these techniques to the genome of H. influenzae by reference to the EcoCyc knowledge base to predict which of 81 metabolic pathways of E. coli are found in H. influenzae. The resulting prediction is a complex hypothesis that is presented in computer form as HinCyc: an electronic encyclopedia of the genes and metabolic pathways of H. influenzae. HinCyc connects the predicted genes, enzymes, enzyme-catalyzed reactions, and biochemical pathways in a WWW-accessible knowledge base to allow scientists to explore this complex hypothesis.

S. M. Paley and P. D. **Karp**, "Adapting CLIM app," ... in *Proceedings of the Association of Lisp Users Meeting and Workshop*, pp. 1--9, 1995.

Abstract: The World Wide Web (WWW) offers the potential to deliver specialized information to an audience of unprecedented size. Along with this exciting new opportunity, however, comes a challenge for software developers: instead of rewriting our software applications to operate over the WWW, how can we maximize software reuse by retrofitting existing applications? We have developed a Web server tool, written in Common Lisp, that allows any existing graphical user interface application written using the Common Lisp Interface Manager (CLIM) to hook easily into the WWW. This tool --- CWEST (CLIM-WEb Server Tool, pronounced "quest") --- has been developed to operate with EcoCyc, an electronic encyclopedia of genes and metabolism of the bacterium E. coli. EcoCyc consists of a database of objects relevant to E. coli biochemistry and a sophisticated interface, implemented in CLIM, that runs on the local host window system and generates graphical displays appropriate to each type of object. Each query to our server is passed as a command to the EcoCyc program, which responds by dynamically generating an appropriate local drawing. That drawing, which can be a mixture of text and graphics, is then translated into the HyperText Markup Language (HTML) and/or the Graphics Interchange Format (GIF) and returned to the client. Sensitive regions embedded in the CLIM drawing are converted to hyperlinks with Universal Resource Locators (URLs) that generate further EcoCyc queries. This tight coupling of CLIM output with Web output makes CLIM an ideal high-level programming tool for Web applications. The flexibility of Common Lisp and CLIM made implementation of the server tool surprisingly easy, requiring few changes to the existing EcoCyc program. The results can be seen at URL http://www.ai.sri.com/ecocyc/browser.html. We plan to make CWEST available to the CLIM community at large, with the hope that it will spur other software developers to make their CLIM applications available over the WWW.

P. D. **Karp**, K. Myers, and T. Gruber, "The generic frame protocol," in *Proceedings of the 1995 International Joint Conference on Artificial Intelligence*, pp. 768--774, 1995.

Abstract: The Generic Frame Protocol (GFP) is an application program interface for accessing knowledge bases stored in frame knowledge representation systems (FRSs). GFP provides a uniform model of FRSs based on a common conceptualization of frames, slots, facets, and inheritance. GFP consists of a set of Common Lisp functions that provide a generic interface to underlying FRSs. This interface isolates an application from many of the idiosyncrasies of specific FRS software and enables the development of generic tools (e.g.,

graphical browsers, frame editors) that operate on many FRSs. To date, GFP has been used as an interface to Loom, Ontolingua, Theo, and Sipe.

Keywords: knowledge representation

P. **Karp**. "A strategy for database interoperation." *Journal of Computational Biology*, vol. 2, no. 4, pp. 573--586, 1995.

Abstract: To realize the full potential of biological databases (DBs) requires more than the interactive, hypertext flavor of database interoperation that is now so popular in the bioinformatics community. Interoperation based on declarative queries to multiple network-accessible databases will support analyses and investigations that are orders of magnitude faster and more powerful than what can be accomplished through interactive navigation. I present a vision of the capabilities that a query-based interoperation infrastructure should provide, and identify assumptions underlying, and requirements of, this vision. I then propose an architecture for query-based interoperation that includes a number of novel components of an information infrastructure for molecular biology. These components include: a knowledge base that describes relationships among the conceptualizations used in different biological databases; a module that can determine the DBs that are relevant to a particular query; a module that can translate a query and its results from one conceptualization to another; a collection of DB drivers that provide uniform physical access to different database management systems; a suite of translators that can interconvert among different database schema languages; and a database that describes the network location and access methods for biological databases. A number of the components are translators that bridge the heterogeneities that exist between biological DBs at several different levels, including the conceptual level, the data model, the query language, and data formats.

P. D. **Karp** and S. M. Paley. "Knowledge representation in the large." in *Proceedings of the 1995 International Joint Conference on Artificial Intelligence*, pp. 751--758, 1995.

Abstract: Frame knowledge representation systems lack two important capabilities that prevent them from scaling up to large applications: they do not support fast access to large knowledge bases (KBs), nor do they provide concurrent multiuser access to shared KBs. We describe the design and implementation of a storage subsystem that submerges a database management system (DBMS) within a knowledge representation system. The storage subsystem incrementally loads referenced frames from the DBMS, and can save only those frames that have been updated in a given session to the DBMS. We present experimental results that show our approach to be an improvement over the use of flat files, and that evaluate several variations of our approach.

Keywords: knowledge representation

P. **Karp** and M. Mavrovouniotis. "Representing, analyzing, and synthesizing biochemical pathways." *IEEE Expert*, vol. 9, no. 2, pp. 11--21, 1994.

Abstract: Living cells are complex systems whose growth and existence depends on thousands of biochemical reactions. A subset of these reactions -- the metabolism -- interconverts small molecules. A variety of computational problems arise in representing knowledge of the metabolism in electronic form, in analyzing that knowledge to gain deeper

insights into complexities of the metabolism, and in using such knowledge in biology, biotechnology and health applications. These problems provide a rich set of opportunities for exploiting existing AI techniques, and challenges for developing new and improved techniques. This article describes challenges and opportunities for addressing computational problems in the metabolism with techniques from knowledge representation, planning, integration of heterogeneous databases, qualitative reasoning, knowledge acquisition, and machine learning. The computational problems include construction of large shared knowledge bases of biochemical pathways, knowledge acquisition from the biochemical literature, qualitative simulation of metabolic pathways, thermodynamic estimation, synthesis of metabolic pathways, and scientific hypothesis formation.

Annotation: This online version was edited heavily to produce the version published in IEEE Expert.

P. **Karp** and S. M. Paley, "Automated drawing of metabolic pathways," in *Third International Conference on Bioinformatics and Genome Research* (H. Lim, C. Cantor, and R. Robbins, eds.), 1994.

Abstract: The EcoCyc system consists of a knowledge base that describes the genes and intermediary metabolism of E. coli, and a graphical user interface (GUI) for accessing that knowledge. This paper presents algorithms for drawing metabolic pathways by dynamically querying the underlying knowledge base. These algorithms provide a foundation for building graphical user interfaces to metabolic databases. Pathway drawing is a graph-layout problem. Our algorithms draw pathways of several different topologies, including linear, cyclic, and branching pathways, as well as larger groupings of such pathways. The algorithms provide several visual presentations of metabolic pathways, for example, compounds can be drawn as names and/or chemical structures, and enzyme names and side compounds can be drawn or omitted. The GUI also provides several facilities for navigating in the space of biochemical pathways, such as traversing connections between pathways, and exploding or collapsing a pathway to include or exclude neighboring pathways.

Keywords: bioinformatics, metabolism

P. D. **Karp**, J. D. Lowrance, T. M. Strat, and D. E. Wilkins, "The Grasper-CL graph management system," *LISP and Symbolic Computation*, vol. 7, pp. 245--282, 1994.

Abstract: Graphs are virtually ubiquitous in programming applications. Moreover, graph-structured information is especially prevalent in AI applications. We can enhance programs that manipulate graph-structured information by providing these programs with graphical user interfaces that draw graphs, and that allow users to interact with drawings of graph nodes and edges. Grasper-CL is a Common Lisp system for manipulating and displaying graphs. Grasper-CL defines a graph abstract datatype and an extensive set of associated operations for creating, modifying and interrogating graphs, and for saving them persistently. The system draws graphs using CLIM (the Common Lisp Interface Manager), and can create postscript renditions of its drawings. Grasper-CL supports a wide variety of graphic styles for drawing graph nodes and edges. The system includes several different automatic graph layout algorithms, such as for circular and tree layout; it also supports full interactive manipulation of graph drawings. Finally, the system provides facilities for building graph-based user interfaces for application programs, which have been used in conjunction with the Sipe planner, the Gister evidential reasoner, a scheduler for the Hubble Space Telescope, and the EcoCyc encyclopedia of biochemical pathways. A number of

groups within the AIC and SRI are using the Grasper-CL system in a variety of projects. This talk will describe the system in detail for people who wish to understand its capabilities better or who are thinking of using it for other projects. This talk is also an opportunity for the audience to shape the future directions of the system: What additional capabilities should be added? Would users like more direct input in how the system evolves? Should we attempt to find funding for further development of the system and research on such issues as graph layout algorithms?

Keywords: graphs

P. D. **Karp**, S. M. Paley, and I. Greenberg. "A storage system for scalable knowledge representation," in *Proceedings of the Third International Conference on Information and Knowledge Management* (N. Adam, ed.), 1994.

Abstract: Twenty years of AI research in knowledge representation has produced frame knowledge representation systems (FRSs) that incorporate a number of important advances. However, FRSs lack two important capabilities that prevent them from scaling up to realistic applications: they cannot provide high-speed access to large knowledge bases (KBs), and they do not support shared, concurrent KB access by multiple users. Our research investigates the hypothesis that one can employ an existing database management system (DBMS) as a storage subsystem for an FRS, to provide high-speed access to large, shared KBs. We describe the design and implementation of a general storage system that incrementally loads referenced frames from a DBMS, and saves modified frames back to the DBMS, for two different FRSs: LOOM and THEO. We also present experimental results showing that the performance of our prototype storage subsystem exceeds that of flat files for simulated applications that reference or update up to one third of the frames from a large LOOM KB.

Keywords: knowledge representation

P. **Karp** and S. M. Paley, "Representations of metabolic knowledge: Pathways," in *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology* (R. Altman, D. Brutlag, P. **Karp**, R. Lathrop, and D. Searls, eds.), (Menlo Park, CA), AAAI Press, 1994.

Abstract: The automatic generation of drawings of metabolic pathways is a challenging problem that depends intimately on exactly what information has been recorded for each pathway, and on how that information is encoded. The chief contributions of the paper are a minimized representation for biochemical pathways called the predecessor list, and inference procedures for converting the predecessor list into a pathway-graph representation that can serve as input to a pathway-drawing algorithm. The predecessor list has several advantages over the pathway graph, including its compactness and its lack of redundancy. The conversion between the two representations can be formulated as both a constraint-satisfaction problem and a logical inference problem, whose goal is to assign directions to reactions, and to determine which are the main chemical compounds in the reaction. We describe a set of production rules that solves this inference problem. We also present heuristics for inferring whether the exterior compounds that are substrates of reactions at the periphery of a pathway are side or main compounds. These techniques were evaluated on 18 metabolic pathways from the EcoCyc knowledge base.

Keywords: bioinformatics, metabolism

P. D. **Karp**. "Design methods for scientific hypothesis formation and their application to molecular biology." *Machine Learning*, vol. 12, pp. 89--116, 1993.

Abstract: Hypothesis-formation problems occur when the outcome of an experiment as predicted by a scientific theory does not match the outcome observed by a scientist. The problem is to modify the theory, and/or the scientist's conception of the initial conditions of the experiment, such that the prediction agrees with the observation. I treat hypothesis formation as a design problem. A program called HypGene designs hypotheses by reasoning backward from its goal of eliminating the difference between prediction and observation. This prediction error is eliminated by design operators that are applied by a planning system. The synthetic, goal-directed application of these operators should prove more efficient than past generate-and-test approaches to hypothesis generation. HypGene uses heuristic search to guide a generator that is focused on the errors in a prediction. The advantages of the design approach to hypothesis-formation over the generate-and-test approach are analogous to the advantages of dependency-directed backtracking over chronological backtracking. These hypothesis-formation methods were developed in the context of a historical study of a scientific research program in molecular biology. This paper describes in detail the results of applying the HypGene program to several hypothesis-formation problems identified in the historical study. HypGene found most of the same solutions as did the biologists, which demonstrates that it is capable of solving complex, real-world hypothesis-formation problems.

P. **Karp** and M. Riley, "Representations of metabolic knowledge." in *Proceedings of the First International Conference on Intelligent Systems for Molecular Biology* (L. Hunter, D. Searls, and J. Shavlik, eds.), (Menlo Park, CA), pp. 207--215, AAAI Press, 1993.

Abstract: Construction of electronic repositories of metabolic information is an increasingly active area of research. Encoding detailed knowledge of a complex biological domain requires finely honed representations. We survey representations used for several metabolic databases, including EcoCyc, and reach the following conclusions. Representation of the metabolism must distinguish enzyme classes from individual enzymes, because there is not a one-to-one mapping from enzymes to the reactions they catalyze. Individual enzymes must be represented explicitly as proteins, e.g., by encoding their subunit structure. The species variation of metabolism must be represented. So must the substrate specificity of enzymes, which may be treated in several ways.

Keywords: bioinformatics, metabolism

P. D. **Karp**. "A qualitative biochemistry and its application to the regulation of the tryptophan operon." in *Artificial Intelligence and Molecular Biology* (L. Hunter, ed.), Menlo Park, CA: AAAI Press, 1993.

Keywords: bioinformatics

P. D. **Karp**. "A knowledge base of the chemical compounds of intermediary metabolism." *CABIOS*, vol. 8, no. 4, pp. 347--357, 1992.

Abstract: This paper describes a publicly available knowledge base of the chemical compounds involved in intermediary metabolism. We consider the motivations for constructing a knowledge base of metabolic compounds, the methodology by which it was

constructed. and the information that it currently contains. Currently the knowledge base describes 952 compounds. listing for each: synonyms for its name. a systematic name. CAS registry number. chemical formula. molecular weight. chemical structure. and two-dimensional display coordinates for the structure. The Compound Knowledge Base (CompoundKB) illustrates several methodological principles that should guide the development of biological knowledge bases. I argue that biological datasets should be made available in multiple representations to increase their accessibility to end users. and I present multiple representations of the CompoundKB (knowledge base. relational data base. and ASN.1 representations) I also analyze the general characteristics of these representations to provide an understanding of their relative advantages and disadvantages. Another principle is that the error rate of biological data bases should be estimated and documented -- this analysis is performed for the CompoundKB.

Keywords: bioinformatics. metabolism

Annotation: The online version has been updated to reflect modifications to the knowledge base.

P. D. **Karp**. "The design space of frame knowledge representation systems." Tech. Rep. 520. SRI International Artificial Intelligence Center. 1992.

Abstract: In the past 20 years. AI researchers in knowledge representation (KR) have implemented over 50 frame knowledge representation systems (FRSs). KR researchers have explored a large and surprisingly diverse space of alternative FRS designs. This paper surveys the FRS design space in search of design principles for FRSs. The FRS design space is defined by the set of alternative features and capabilities -- such as the representational constructs and inheritance mechanisms -- that an FRS designer might choose to include in a particular FRS. as well as the alternative implementations that might exist for a particular feature. The paper surveys the architectural variations explored by different system designers for the frame. the slot. the knowledge base. for the inheritance operation. and for access-oriented programming and object-oriented programming. We also discuss the classification operation in detail. We find that few design principles exist to guide an FRS designer as to how particular design decisions will affect qualities of the resulting FRS. such as its worst-case and average-case theoretical complexity. its actual performance on real-world problems. the expressiveness and succinctness of the representation language. the runtime flexibility of the FRS. the modularity of the FRS. and the effort required to implement the FRS.

Keywords: knowledge representation

P. D. **Karp**. "Artificial intelligence methods for theory representation and hypothesis formation." *CABIOS*. vol. 7. no. 3. pp. 301--308. 1991.

Keywords: machine learning

V. K. Chaudhri. A. Farquhar. R. Fikes. P. D. **Karp**. and J. P. Rice. "The Generic Frame Protocol 2.0." *Artificial Intelligence Center. SRI International*. 21 July 1997.